

Evaluating the Diagnostic Validity of a Facet-based Formative Assessment System

Angela Haydel DeBarger¹, Louis DiBello², Jim Minstrell³, Mingyu Feng¹, William Stout²,
James Pellegrino², Geneva Haertel¹, Christopher Harris¹, and Liliana Ructinger¹

¹ Center for Technology in Learning, SRI International; ² University of Illinois at Chicago;
³ FACET Innovations, LLC

**Paper presented at the Society for Research on Educational Effectiveness Fall Conference,
Washington, DC, September 2011.**

Corresponding Author:

Angela Haydel DeBarger, Center for Technology in Learning, SRI International, 333 Ravenswood Avenue,
Mailstop BN335, Menlo Park, CA 94025. Tel. 650-859-2433. Email: angela.haydel@sri.com.

Acknowledgments:

This material is based upon work supported by the U.S. Department of Education under Grant Number R305A100475.

ABSTRACT

This paper describes methods for an alignment study and psychometric analyses of a formative assessment system, Diagnoser Tools for physics. Diagnoser Tools begin with facet clusters as the interpretive framework for designing questions and instructional activities. Thus each question in the diagnostic assessments includes distractors that represent common problematic ideas in physics. Although Diagnoser Tools appear to have instructional validity, current alignment and psychometric approaches have not been sufficient for validating the complex, multidimensionality of the facets and facet-based diagnostic questions. Together the recent alignment and psychometric studies provide a more comprehensive picture of the validity of the facet-based assessment system.

Evaluating the Diagnostic Validity of a Facet-based Formative Assessment System

Facet-based assessments are one innovative approach to helping teachers diagnose students' science understanding (Minstrell, 2001; Minstrell, Anderson, Kraus, & Minstrell, 2008). The facets perspective assumes that students' understandings possess some strengths to build on, possibly in addition to problematic thinking that can be revised through additional learning opportunities. The term "facets" acknowledges that not all students' thinking can be considered "misconceptions" or errors. Facet clusters serve as the interpretive framework for analyzing student responses to questions and for designing instructional activities to promote learning in the online Diagnoser Tools formative assessment system (www.diagnoser.com).

Despite the designed-in multidimensionality of facet-based assessments, the preponderance of psychometric analyses performed so far have failed to capture the richness of the evidence about what students know and how they know it. Standard classical test theory or unidimensional approaches can be useful for capturing some critical measurement properties of items and instruments [such as gross item indices of "difficulty," biserial correlations with total score, classical unidimensional indices of total score reliability including KR20 and Cronbach's alpha (Cronbach, 1951; Lord & Novick, 1968; Allen & Yen, 2002; van der Linden & Hambleton, 1997), and unidimensional IRT analyses], but are not designed to reflect the facet-based multidimensional richness of the data.

In light of the rich conceptual and cognitive models guiding item development and data collection, the failure to use more powerful measurement models means that the linkage from observation to interpretation and from interpretation back to cognition is only of the most rudimentary form. The work reported here provides a strong interpretive framework supported by sophisticated psychometric techniques as a way of capturing the diagnostic power of the instrument, and enhancing its usefulness as a formative assessment tool. Our approach provides insight about the potential of the facet-based approach to offer a clear and transparent articulation of the linkage between the assumptions about cognition and observed student performance.

Components of the research project include: (a) an Evidence-Centered Design analysis of facet-based instructional materials and assessments that provides a view of the evidentiary coherence of the existing system; (b) an alignment study of the Diagnoser system with multiple standards frameworks that describes deep connections among existing standards frameworks and the Diagnoser system and illuminates how alignment approaches can simultaneously inform all aspects of a formative assessment system; (c) the application of sophisticated psychometric models to the existing data that provides statistical evidence for inferential claims that support classroom use of the Diagnoser system; and (d) the identification of cases in which new or improved Diagnoser question sets can be developed and tested with students.

This paper focuses on methodologies associated with the psychometric analyses of the Diagnoser question sets and the alignment study of facets and question sets. Force and Motion content was targeted in this project for two reasons. First, extensive research has been conducted on misconceptions in force and motion (e.g., Driver et al., 1994; Hashweh, 1988; Lythcott, 1985; Wandersee et al., 1994) and by staff of several related Diagnoser projects, providing a substantial research basis for the design of the instructional materials and assessments. Second, Force and Motion clusters are among the earliest developed as part of the online Physics Diagnoser and have been used continually since 2004 during which has been available via the internet, providing a large database of student performances for analysis.

Diagnoser Tools

Diagnoser Tools is an online system developed by Facet Innovations to support teachers in enacting the practice of formative assessment in the subject fields of middle school and high school physics, physical science and chemistry. Effective formative assessment practice requires that teachers and students establish a common understanding of learning goals, teachers elicit student thinking, teachers respond to and make sense of student thinking in relation to learning goals, teachers take action based on their interpretation of students' learning needs, and teachers re-assess (Black and Wiliam, 2009). Effective elicitation of student ideas often involves dialogic conversation, particularly to identify problematic ways of reasoning. Diagnostic assessments also can be used to give teachers a “snap-shot” of conceptions in the class.

Diagnoser Tools includes questions to elicit and engage students in conversation about their ideas (elicitation questions), lessons that prompt students to explore their ideas (developmental lessons), diagnostic questions (Diagnoser question sets), and activities targeted to address specific problematic ideas (prescriptive activities). Each component references a facet cluster. In this sense, facet clusters serve as the backbone of the system. Our early work in this project focuses primarily on two components of this system—the facet clusters and the Diagnoser question sets.

Facet Clusters

Each facet cluster contains learning goals and common problematic student ideas related to these learning goals. Figure 1 shows the facet cluster for Forces as Interactions. Facets are arranged with the goal facets at the top followed by more problematic facets. The 0X and 1X facets are the goal facets. The 2X through 9X facets indicate ideas that have some problematic aspects. In general, the higher number facets (i.e., 9X, 8X) are the most problematic. The X0s indicate more general statements of student ideas and are sometimes followed by specific examples, coded as X1 through X9.

Figure 1. Forces as Interactions Facet Cluster

00	The student understands that all forces arise out of an interaction between two objects and that these forces are equal in magnitude and opposite in direction.
01	All forces arise out of an interaction between two objects.
02	The force pairs are equal in magnitude.
03	The force pairs are opposite in direction.
<hr/>	
40	The student identifies equal force pairs, but indicates that both forces act on the same object. (For the example of a book at rest on a table, the gravitational force down on the book and the normal force up by the table on the book are identified as an action-reaction pair.)
50	The student uses the effects of a force as an indication of the relative magnitudes of the forces in an interaction.
51	More damage indicates one of the interacting objects exerted a larger force.
52	If an object is at rest, the interaction forces must be balanced.
53	If an object moves, the interaction forces must be unbalanced.
54	If an object accelerates, the interaction forces must be unbalanced.
60	The student indicates that the forces in a force pair do not have equal magnitude because the objects are dissimilar in some property (e.g., bigger, stronger, faster).
61	The 'stronger' object exerts a greater force.
62	The moving object or a faster moving object exerts a greater force.
63	The more active or energetic object exerts more force.
64	The bigger or heavier object exerts more force.
90	The student believes that inanimate/passive objects cannot exert a force.

Question Sets

One or two sets of questions are associated with each facet cluster. As a student works through a question set, she receives targeted feedback after responding to each item. Most response options for selected-response questions are linked to a particular facet, and the feedback and subsequent question shown to the student are based on the facet that was diagnosed from the response to the previous question. Figure 2 shows two items from a question set on Forces as Interactions and a table showing the facet code assigned to each answer option of Question 1.


Figure 2. Questions from Forces as Interactions Diagnoser Question Set

Question: 1

Jennifer and Katie stand and lean on each other.

Jennifer weighs 150 pounds and Katie weighs 120 pounds.

Which one pushes harder on the other?



☐ [a] Katie must push harder because she weighs less and has to compensate for having less weight.
 ☐ [b] Jennifer and Katie push on each other with the same size force because force pairs are always equal.
 ☐ [c] Jennifer pushes harder because she weighs more.
 ☐ [d] It depends on whether Jennifer or Katie moves.

Key	Facet
a	63
b	02
c	64
d	50

Question: 2

In the space below, describe who exerts the greater force in each of the following conditions AND why.

1. If Katie moves
2. If Jennifer moves
3. If neither moves

Facet-based assessments have been designed in such a way that most multiple-choice distractors are linked to particular facets of understanding that represent complete, partial or incorrect understanding. For example Question 1 states: “Jennifer and Katie stand and lean on each other. Jennifer weighs 150 pounds and Katie weighs 120 pounds. Which one pushes harder on the other?” If a student chooses Option C: “Jennifer pushes harder because she weighs more,” then Diagnoser reports that this response indicates that the student’s way of thinking corresponds to Facet 64, “The bigger or heavier object exerts more force.”. From Figure 1, we see that Facet 64 falls within a more general problematic understanding: Facet 60, “The student indicates that the forces in a force pair do not have equal magnitude because the objects are dissimilar in some property (e.g. bigger, stronger, faster).” Question 2 follows Question 1 and requests students to show their reasoning in addition to answering the multiple choice question.

Some questions are composed as a related pair of questions, where the second question of the pair asks the student to provide a reason for the response given to the first question. For these paired questions, a facet is inferred from responses to both questions. An example can be found in Questions 6 and 7, also in Figure 3. For instance, if a student chose Option a for Question 6 and Option b for Question 7, the system will diagnose the student thinking as Facet 53.

Figure 3. Paired Questions in Forces as Interactions Diagnoser Question Set

Question: 6

Sarah plays defensive back on her school's soccer team. At practice she kicks the ball that was rolling toward her to the other end of the field.

Which statement describes the force by the ball acting on Sarah's foot **during the kick**?

☐ [a] The ball does not exert a force on Sarah's foot.
☐ [b] The force by the ball is less than the force of Sarah's kick.
☐ [c] The force by the ball is equal to the force of Sarah's kick.
☐ [d] The force by the ball is greater than the force of Sarah's kick.

Key	Facet
a	Paired with Question: 7
b	Paired with Question: 7
c	Paired with Question: 7
d	Paired with Question: 7

Question: 7

Paired with Question: 6

Which reason best fits your answer to the previous question?

☐ [a] Sarah is stronger than the ball.
☐ [b] Sarah's kick made the ball move, but the ball did not move Sarah.
☐ [c] Only Sarah can exert a force; the ball is not alive.
☐ [d] All interacting objects exert equal forces on each other.
☐ [e] The ball hurt Sarah's foot more than she hurt the ball.
☐ [f] The ball was moving when Sarah kicked the ball.

First Key	Key	First Facet	Facet
a	a	90	61
	b	90	53
	c	90	90
	d	90	01

Note. Only facet combinations with Question 6 Option a combinations with Question 7 are shown.

Teachers receive real-time results of students' answers in a report describing the facets used by each student. The report also contains summary statistics for the percentages of the class responding with particular facets. Students also receive a report on their performance that highlights some things they may need to work on. A teacher's report consists of a list of the inferred facets, one to each question or related pair of questions. As can be seen in Figure 4, students' responses varied in degree of internal consistency across the questions. For example, Student 6, who responded with Facet 64 to Question 1, was therefore branched to Question 4, followed by Questions 5, 6 and 7, and facets associated with her remaining responses were 00, 02, 02, 01. Since facets 0x represent learning targets, this student, after the first response, demonstrated consistent goal knowledge and understanding. By contrast, the response facets for Student 1 are 50, 40, 54, 90, 60, 62, 51; a very inconsistent indication of any specific facet of understanding for this student.

Figure 4. Forces as Interactions Teacher Report

Forces as Interactions Set 1			Questions							
Student ID	Date Completed	Self Rating	1	2 All	3	4	5	6	7	8 All
1239-0-1	2005-10-22	5	50	text		40	54	90 , 60	62 , 51	text
1239-0-2	not yet	none	63			00	02	02	01	
1239-0-3	2005-10-22	3	02			53	02	02	01	
1239-0-4	2005-10-22	3	50	text		00	53	02	01	
1239-0-5	2005-10-22	2	63			00	02	53 , 02	53 , 01	
1239-0-6	2005-10-22	2	64			00	02	02	01	
1239-0-7	2005-10-22	2	02			00	02	53 , 02	53 , 01	
1239-0-8	2005-10-22	5	64			00	53	02	01	
1239-0-12	2005-10-22	2	02			00	02	02	01	
1239-0-13	2005-10-22	3	02			63	53	02	01	
1239-0-14	2005-10-22	6	02			63	02	02	01	
1239-0-15	2005-10-22	2	50	text		00	02	53 , 02	53 , 01	
1239-0-16	2005-10-22	4	02			63	64	Unk , 02	62 , 01	

Summary Statistics for Forces as Interactions Set 1

Facet	At least once	More than Once
00	92.3%	92.3%
40	7.7%	0%
50	61.5%	38.5%
60	61.5%	15.4%
90	7.7%	0%
Unk	7.7%	0%

[Explanation for Facets and notes](#), [Text of questions](#), [Prescriptive activities](#)

Instructional and Diagnostic Validity of Diagnoser Question Sets

A fundamental assumption that applies to facets-based assessments is that the items tap a range of critical understandings and cognitive skills in the content domain. In essence, the assessment instrument and its items are inherently multidimensional by design. Such multidimensionality is a potential strength that can be captured for the productive purposes of profiling facets of students' understanding and generating interpretive results that instructors could use in their teaching. Responses to individual items provide information about likely student misconceptions or partial conceptions, and *patterns* of responses to a set of facet-based questions are expected to provide strong evidence about a student's knowledge and understanding. This work aims to develop and examine formal methods that can psychometrically and statistically support the typical diagnostic uses of the Diagnoser Teacher Reports as illustrated in Figure 4.

Facet-based assessments appear to have instructional validity, but not much research has been done on issues of interpretation of facet-based assessments. Without such research the interpretability and use of these assessments is severely restricted. Studies that validate the conceptions represented as well as the relation between facets and items in question sets have never been conducted. In addition, psychometric modeling and analysis typically performed with these data have fallen short of reflecting the true design and purpose of the assessments (Scalise, Madhyastha, Minstrell, and Wilson, 2010; Steedle & Shavelson, 2009; Steedle, 2008; Wilson 1992, 2008).

Our approach for examining the diagnostic validity of Diagnoser Tools takes into account that Diagnoser Tools function as a system, rather than as discrete components. Facets within a cluster are intended to relate to one another in meaningful ways. Each of the components (i.e., lessons, question sets) are intended to cohere within each facet cluster. Moreover, interpretation of evidence about the validity of these components must take account of when lessons and question sets are intended to be used in instruction. Although traditional alignment and psychometric analyses may provide some information about the validity of these components, they are not sufficient. Our research questions and approach reflect this perspective.

Alignment Study

In contrast to typical alignment studies, this study will provide more substantial evidence about the diagnostic capabilities of the Diagnoser formative assessment system. Because facet clusters are a foundational piece to which other components of Diagnoser Tools relate, several research questions for the alignment study focus on the content, diagnostic, and cognitive validity of the facets in the cluster, as shown in Table 1. We will examine goal facets in relation to standards (research question 1). Because problematic facets are essential for designing diagnostic questions, we also attend to the representation and ordering of problematic facets (research questions 2 and 3).

Table 1. Research Questions and Approach for Alignment Study of Facet Clusters

Research Questions	Alignment to Internal or External Criteria or Other	Type of Validity	Rationale
RQ1: To what degree do standards and facet clusters address the same content categories?	External: Alignment of goal facets to 3 standards frameworks (Benchmarks for Scientific Literacy, College Board Standards objectives, and New Framework for Science Education core and component ideas)	Content validity	Confirmation of existing alignment to BSL; Documentation of range of knowledge addressed in relation to new frameworks
RQ2: Do problematic facets represent the range of frequent 'misconceptions' and problematic ways of thinking?	External: Alignment of problematic facets to misconceptions reflected in research	Diagnostic validity; Cognitive validity	Confirmation that problematic facets are representative of student cognition in physics and can provide relevant diagnostic information
RQ3: Does the ordering of problematic facets appropriately reflect less to more problematic ideas?	Other: Judgment of the degree to which the ranking of the problematic facts reflect an order of less to more problematical	Diagnostic validity; Cognitive validity	Informs extent to which clusters appropriately represent PFs as more or less problematic—this would affect teacher's perceptions of the extent to which students have significant difficulty with the physics content. Informs understandings about the relationships of misconceptions and problematic ideas in students' thinking about physics. Not necessarily evidence of whether facet clusters are a learning trajectory.

The analysis of question sets also is designed to evaluate content, diagnostic, and cognitive validity, as shown in Table 2.

Table 2. Research Questions and Approach for Alignment Study of Diagnoser Question Sets

Research Questions	Alignment to Internal or External Criteria or Other	Type of Validity	Rationale
RQ4: To what degree do Diagnoser questions align to facet clusters? (To what degree does the correct answer align to the goal facet? To what degree do the incorrect answers align to the problematic facets?)	Internal: Alignment of each question to facet cluster	Diagnostic validity; Indirect content validity	Confirmation that questions are likely to elicit intended facets. Indirect content alignment signifies that if questions are aligned to facet clusters and facet clusters are aligned to content standards, we can infer that questions are aligned to content standards.
RQ5: What degree of depth or complexity of knowledge do Diagnoser questions address?	External: Alignment of each question to components reflecting depth of knowledge (declarative, procedural, schematic, or strategic)	Cognitive validity	Evidence about the complexity of questions and sets
RQ6: How well do the reports communicate student performance in ways that teachers understand and inform next instructional steps?	Other: Judgment of the effectiveness of the Diagnoser question set reporting functionality for supporting teachers' formative use of the question sets	Diagnostic validity	Evidence about effectiveness of Diagnoser data for teacher use
RQ7: To what extent do the pathways through the questions in each Diagnoser Question Set represent a logical, conceptually appropriate sequence?	Other: Judgments about whether use of paired questions, skips and repeats is logical and appropriate for diagnosing student thinking; Is question order logical (e.g., in terms of when facets are elicited)?; What do common response pathways reveal about the diagnostic capabilities of the question sets?	Diagnostic validity	Analysis will provide insight into similarities and differences in question set design and flow, and perhaps also differences in the diagnostic capabilities of question sets.

Method

Identification and Distribution of Panelists. Six experts in physics content, physics instruction, science standards, and cognitive science will be selected as panelists. Panelists will be divided into two

groups. We will assign alignment tasks so that group effort is approximately equivalent. Group 1 will consist of three panelists with expertise in physics content, physics instruction and science standards. Group 1 panelists will respond to research questions 1 and 4. Group 2 will consist of three panelists with expertise in physics content, physics instruction and cognitive science. Group 2 panelists will respond to research questions 2, 3, 5, 6 and 7. Although Group 2 panelists will be responding to more research questions, we expect that the amount of time to complete judgments will be similar to the time needed for Group 1 because of the difference in the nature of the research questions.

Panelists first will convene via teleconference to learn about the Diagnoser system and receive training on how to complete the alignment for research questions 1, 2, and 3. Panelists will submit their responses online. Panelists will convene at a common location to receive training on how to complete the judgments for the remaining questions. Training and completion of ratings will occur over 2 days.

Judgments of Facet Clusters. Seventeen facet clusters related to the three key concept strands in Force and Motion: 1) Description of Motion, 2) Nature of Forces, and 3) Forces to Explain Motion will be included in the study. Panelists will judge each facet cluster's alignment to two nationally recognized frameworks: *AAAS Benchmarks for Scientific Literacy* (AAAS, 1993) and *Science: The College Board Standards for College Success* (College Board, 2009). When standards based on the new framework for K-12 science education: *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (NRC, 2011) are available, we will reconvene panelists to review facet clusters in relation to these standards. For research question 1, the panelists will make a judgment about the degree to which the goal facets and sub-goal facets in each cluster align with the standards using a 4-point Likert scale (1 = Not aligned to 4 = Fully aligned; "Fully aligned" means that the set of facets addresses all components of the associated standards and "Not aligned" means that the set of facets addresses none of the components of the associated standards.) For research question 2, panelists also will judge the extent to which the problematic facets in each cluster reflect common misconceptions and problematic ways of thinking documented in the K-12 literature associated with the topic within the cluster related, using a 5-point Likert scale (1 = Very poor coverage to 5 = Very good coverage). For research question 3, panelists will consider whether the ordering of the problematic facets reflects how these problematic ways of thinking are in learning physics and whether ideas that represent more significant learning needs are more likely to prevent students from attaining the learning goal are given a higher facet code. The panelists will then use a 3-point Likert scale (1 = not at all appropriate to 3 = very appropriate) to rate the degree to which the ranking of the problematic facets in Diagnoser reflect an order of less to more problematical.

Judgments of Diagnoser Question Sets. Of the 17 facet clusters included in the study, most have two associated Diagnoser Question Sets. Each question set typically has 7-12 questions. A total of 32 Diagnoser Question Sets will be included in the study. For each item in the Diagnoser question sets, panelists will evaluate the extent to which the response options exemplify the facets as they are coded, using a 3-point Likert scale (1=No, 2=Somewhat, 3=Yes). Panelists will also identify the particular types of knowledge, declarative, procedural, schematic, or strategic (Li, Shavelson & White, 2002; Ruiz-Primo, 2002; Ruiz-Primo et al., 2002, Shavelson & Ruiz-Primo, 1999) that students may use to answer each question by reviewing a low inference checklist. Because the data from question sets are critical for helping teachers make instructional decisions, panelists will indicate how effectively the reports: (a) communicate what students know and can do and (b) inform next instructional steps using a 4-point rating scales (Not at all effective/useful to Very effective/useful) and open-ended text responses to questions on how the reports can be improved to better communicate what students know and can do or to better inform instruction.

Finally, for each question set panelists will review three common pathways through the question sets (based on actual student data), as not all students complete the same sets of questions or complete a same set of questions in the same sequence. For each pathway panelists will evaluate whether questions are presented in a logical, conceptually appropriate sequence for use by teachers to diagnose and remediate the misconceptions the students hold. Panelists will respond using a 4-point scale ranging from strongly disagree to strongly agree.

For each question in the alignment protocol, all judgments will be made using an online survey system to facilitate analyses. Judgments can easily be exported to examine interrater agreement and to summarize findings for facet clusters and question sets. If possible, discrepant ratings will be resolved by consensus.

Future Analyses of Elicitation Questions, Developmental Lessons and Prescriptive Activities. To complete a comprehensive analysis of the Diagnoser system, a second phase of the alignment study will include an analysis of the elicitation questions, developmental lessons and prescriptive activities. Panelists will confirm the alignment of elicitation questions to facet clusters. Developmental lessons and prescriptive activities will be analyzed not only in terms of alignment to facet clusters, but also to examine how they engage students in the scientific practices (e.g., data interpretation, explanation, modeling),

Psychometric Analyses of Diagnoser Question Sets

Diagnostic Question Set Data. Existing Physics Diagnoser data available for analysis have been collected through the online website <http://www.Diagnoser.com/>. The primary student outcomes, as shown in the Teacher Report, Figure 4, are the “facet scores” from a student’s question set responses, which are a sequence of facets derived from a student’s multiple-choice responses. In addition to the facet codes linked to each multiple-choice response, we include as a further dimension the type of context for a given question or multiple choice option. For example student’s responses to Diagnoser questions appear to be affected by whether acting objects are animate or inanimate, or whether the question situation is a real-world setting or is more abstract. We have developed a list of context types for investigation. We summarize and represent students’ facet scores within context type for a given question set in three ways: (1) prevalence across the question set of inferred facets; which facets occurred and how often relative to types of question context; (2) patterns of inferred facets; and (3) degree of consistency within a given student’s facet pattern.

Advanced psychometric analyses focus on two main objectives: (1) to summarize the observed ranges and patterns of performance of a variety of groups of students in question sets associated with each cluster; and (2) to provide a definitive evaluation of the diagnostic capacity of the question sets with respect to inferring aspects of students’ facet profiles in light of the formative purposes of these assessments. Previous model-based psychometric studies have looked at only a few clusters and facet-based question sets and have had limited success in capturing students’ observed performance on items in ways that were sensitive to the cluster and facets structure (Scalise, Madhyastha, Minstrell, and Wilson, 2010; Steedle & Shavelson, 2009; Steedle, 2008; Wilson 1992, 2008). Here we are analyzing available data from all seventeen clusters and thirty-two question sets, and are employing a novel set of approaches that take into account structural and design characteristics of the facet assessments in light of their classroom diagnostic purpose. For example, highly promising diagnostic model-based analyses have been completed of another type of misconception-based assessment called concept inventories (Santiago-Román, 2009; Santiago-Román et. al., in preparation). In that work a set of “skills” for

diagnostic measurement and reporting was shown to be highly predictive of student performance. We are carrying out similar diagnostic analyses for facet-based assessments, guided by findings from the alignment study and ECD analyses, and we are using a constrained latent class model for cognitive diagnosis called the Fusion Model (DiBello, Roussos & Stout, 2007; DiBello & Stout, 2003; DiBello & Stout 2007). In addition, a new diagnostic model is being applied that goes beyond question right/wrong scoring and takes account of facet information linked to multiple choice options (DiBello, Stout & Henson, in press; for a different approach see de la Torre 2009). We also are performing unidimensional analyses to establish baseline information about performance on individual items and on the whole test. For that purpose we are applying a partial credit model that allows for two or more possible responses to have the same ordered score level to scale the response data (Wilson 1992, 2008).

Classroom Formative Focus. This psychometric and statistical approach to the Physics Diagnoser data takes account of three key aspects of these assessments: (1) the chief purpose being for teachers to identify one or two aspects of thinking that could profitably be discussed with a student or class; (2) incorporation of a number of contextual features of questions and multiple choice options that affect student performance; and (3) the “low stakes” nature of typical formative classroom uses of these facet-based assessments. For example, the inclusion of question context as part of the measurement model acknowledges the reality that students’ question responses are best interpreted in light of the context of the questions or responses. Whether a student’s possession of a particular facet of thinking is robust across multiple contexts is a more difficult question that is not decidable from the Diagnoser question sets, given their short length. For example eight context categories are being investigated, each with multiple context types within category. A given question set with ten or fewer questions on average cannot provide a robust coverage of all or most of the relevant contexts in interaction with the facets of the given cluster. Based on the short lengths of Diagnoser question sets and consideration of multiple contexts, any attempt to infer a student’s facet construction of knowledge will on average have relatively low reliability and will reflect only a small number of relevant context types. The classroom formative focus for the question sets is on identifying aspects of students’ thinking that would be most profitable for informing decisions about next activities, for example for facilitating a classroom discussion. In practical terms, given the context types that were not included in the given question set, even if a particular student’s question set indicates high consistency in selecting question responses that correspond to the goal facet in a given cluster, we may still wish to identify a problematic facet of thinking about which to engage in conversation with that student or group of students.

New Psychometric Paradigm. Given the status of question set results as one of multiple sources of information available to teachers and students in the classroom, a different psychometric paradigm is suggested. The pertinent psychometric question is what is the strength of evidence or information provided by a given Diagnoser question set for the purpose of updating a teacher’s and student’s prior beliefs about students’ constructions of knowledge? Given what the teacher knows about a particular student or class prior to the administration of a given question set, how can the teacher’s prior beliefs about the student or class be modified or updated as a result of the scored Diagnoser question set responses? We adopt the statistical notion of Bayesian updating of prior beliefs, where we mean an application of Bayes Theorem (Bayes, 1763; Stigler, 1982) that allows one to express new beliefs about a student or class based on new evidence and prior beliefs. Bayesian updating takes the place of the classical standardized test notion of an assessment event that is isolated from all other teaching and learning events and that must stand entirely on its own. By contrast, a high school physics classroom typically unfolds as a series of occurrences: discussion, activity, discussion, hands-on activity, discussion, and then perhaps Diagnoser question set. The activities before the question set allow teachers and

students to build sets of prior beliefs about students' constructions of knowledge. The Diagnoser question set provides one new piece of information to students and teachers at a given point of time. The operative question in this situation becomes how effectively does the question set enable teachers and students to update their prior beliefs about students' current constructions of understanding? What are the implications about design of questions and question sets for supporting the updating of beliefs?

Thus, we examine the effectiveness of a Diagnoser question set for providing additional evidence about individual students and groups of students that is psychometrically supported and informative for updating prior beliefs about a student's apparent construction (facet). The traditional canons of high classical unidimensional reliability are moot for Diagnoser questions sets in such a context. An incorrect or inaccurate signal about a student as interpreted from a facet based assessment result very likely will have an opportunity to be corrected during the succeeding few days as the class proceeds, and a functioning classroom will demonstrate a degree of self-correction. Acknowledgement of such a Bayesian updating paradigm within a set of low stakes uses of the facets based assessments shifts the measurement focus to identifying one or two facets of problematic and correct thinking that may be worth attention by the student and teacher in the classroom. Thus, measurement here functions directly in the service of instructional decision making, and an acknowledgement of the partial nature of the information provided by any single question set is part of the effective interpretation of the results as evidence-based promotion of useful classroom discourse.

We are beginning our investigations with groups of students for whom we have data on three closely related facet clusters: Forces as Interactions, Explaining Constant Speed, and Explaining Changes in 1-Dimensional Motion. For these students we have responses to as many as six Question Sets, two for each of the three facet clusters named. We are examining students' facet prevalence values and patterns relative to context features of the questions and how they relate to students' diagnostic model-computed facet profiles. We also are comparing students' performances at middle school and high school levels within and between clusters. To ensure that estimates of standard error, effect sizes, and statistical significance are not artificially inflated or deflated, we are accounting for clustering of students within classes by performing Hierarchical Linear Modeling (Raudenbush & Bryk, 2002) analyses.

Coherence among Alignment and Psychometric Studies

Together the present alignment and psychometric studies provide a more comprehensive picture of the validity of the facet-based assessment system. The studies are complementary, and each informs the other. If as part of the alignment study experts confirm that facet clusters are indeed meaningful groups of physics concepts/ideas, that items are aligned to these facets, and that the question set reports present information to teachers that help them make instructional decisions, we have more confidence in this formative assessment system. The psychometric analyses provide statistical evidence for these inferences about what questions tell teachers about student thinking.

Additionally, components of each study can inform the other. For example, the psychometric studies will take account of judgments made by experts about alignment to various sets of standards. In addition, the alignment study findings may suggest that certain facets are more or less equivalent or that questions intended to address one facet actually appears to relate to different facets. These findings will have implications for the measurement models. The alignment study also can incorporate psychometric findings. For example, in reviewing question set pathways, "common" pathways can be informed by actual student data.

Considerations and Implications

Questions and diagnostic assessments when used as part of formative assessment have a clear purpose of helping teachers elicit, engage and diagnose student thinking so that they can respond to students' learning needs and facilitate sense-making in class discussions and lessons. Thus, they must be evaluated not only in terms of whether they tap appropriate standards, but also in terms of whether they elicit targeted goal understandings and problematic facets of thinking and whether they actually support teachers in making instructional decisions based on student conceptions. Because questions in Diagnoser Tools are based on facet clusters that supposedly represent the core conceptual understandings and problematic ideas, these facet clusters also must be evaluated.

A key contribution of this project is its application of rigorous methods of psychometric and evidentiary analyses to the existing Physics Diagnoser system and its extensive existing database of student performance. The psychometric analyses apply new diagnostic and statistical approaches to a very large database of existing data. These analyses will logically and statistically test the degree of concordance of the existing system and the substantial facets system developmental foundations and design which were based on pragmatic, subject area instructional and pedagogical expertise and early literature on physics teaching and learning and on misconception research.

The Alignment Study implements a new approach for a comprehensive analysis of an assessment system designed for formative use. Traditional alignment approaches typically only examine questions in relation to standards, but they do not articulate a methodology for evaluating the appropriateness of problematic ideas in a framework which is guiding the assessment design. In addition, because we want to evaluate the diagnostic and cognitive validity of questions, we must examine whether items not only elicit goal ideas, but also problematic ideas. Thus, new approaches/types of research questions were needed to investigate the representation of these problematic facets. To evaluate the cognitive demands of items, we also are examining whether items elicit declarative, procedural, schematic, and/or strategic knowledge. Finally, because sets of questions are meaningful units in and of themselves, we ask questions about how the questions in a set work together to support the teacher in making instructional decisions based on data from the question sets.

In concert these approaches are assembling multiple strands of evidence to build a powerful, theoretically and empirically grounded validity argument that directly impacts the pragmatic successes of the Diagnoser Tools system, and that is expected to lead to methods for improving the questions to be tested with students in the second half of the project. In sum, this project applies a rigorous comprehensive approach to understanding the cognitive, instructional and inferential underpinnings of the Diagnoser Tools system in light of the Diagnoser's pragmatic classroom applications.

We believe that our approach is generalizable beyond Diagnoser Tools for physics. The tools and methods developed here certainly can be applied to Diagnoser Tools in other domains (e.g., chemistry and human biology). These approaches to psychometrics and alignment should also work for other comprehensive formative assessment systems that incorporate tools for teachers that provide learning goals and targeted problematic ideas, questions and lessons to elicit and develop student ideas, and questions that are intended to supply diagnostic information related to learning and teaching.

REFERENCES

- Allen, M. J., & Yen, W. (2002). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.
- American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- Bayes, Thomas; Price, Mr. (1763). "An Essay towards solving a Problem in the Doctrine of Chances." *Philosophical Transactions of the Royal Society of London*, 53: 370–418.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, & Accountability*, 21(1), 5-31.
- College Board (2009). *Science college board standards for college success*. New York: Author.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement*, 33, 163-183.
- DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26, psychometrics* (pp. 979-1030). Amsterdam: Elsevier.
- DiBello, L. & Stout, W. F. (2003). Student profile scoring methods for informative assessment. In H. Yanai, A. Okada, K. Shigemasa, & J. J. Meulmann (Eds.) *New developments in psychometrics* (pp. 81-92). Tokyo: Springer.
- DiBello, L.V. & Stout, W. F. (2007) (Invited Guest Co-editors). Special issue on IRT-based cognitive diagnostic models and related methods, *Journal of Educational Measurement*, 44(4), 285-292.
- DiBello, L.V., Stout, W.F., & Henson, R.A. (in press) Cognitive Diagnostic Models for Multiple Choice Assessments with Misconception-Linked Distractors.
- Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas*. London: Routledge.
- Hashweh, M. (1988). Descriptive studies of student's conceptions in science. *Journal of Research in Science Teaching*, 25, 121-134.
- Li, M., Shavelson, R. J., & White, R. T. (2002). *Toward a framework for achievement assessment design: The case of science education*. Stanford CA: School of Education, Stanford University.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley Publishing Company.
- Lythcott, J. (1985). "Aristotelian" was given as the answer, but what was the question? *America Journal of Physics*, 53, 428-432.
- Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications for professional, instructional, and everyday science*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Minstrell, J. A., Anderson, R., Kraus, P., & Minstrell, J. E. (2008). Bridging from practice to research and back: Tools to support formative assessment. In J. Coffey, R. Douglas, & C. Sterns (Eds.), *Science assessment: Research and practical approaches*. Arlington, VA: NSTA Press.

- National Research Council. (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models* (Second Edition). Thousand Oaks: Sage Publications.
- Ruiz-Primo, M.A. (2002, February) On a seamless assessment system. Paper presented to the AAAS annual meeting, Boston, MA.
- Ruiz-Primo, M.A., Shavelson, R.J., Hamilton, L. and Klein, S. (2002) On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39, 369–93.
- Santiago-Román, A. I., DiBello, L., Streveler, R., & Steif, P. (in preparation) Diagnostic capability of concept inventories: A new application of the Fusion Model.
- Santiago-Román, A. I. (2009) Fitting cognitive diagnostic assessment to the concept assessment tool for statics (CATS). Doctoral dissertation, Purdue University.
- Scalise, K., Madhyastha, T., Minstrell, J., & Wilson, M. (2010). Improving assessment evidence in e-learning products: Some solutions for reliability. *International Journal of Learning Technology, Special Issue: Assessment in e-Learning*.
- Shavelson, R.J. and Ruiz-Primo, M.A. (1999) On the assessment of science achievement.(English version) *Unterrichts wissenschaft*, 27(2), 102–27.
- Steedle, J. T. (2008). Latent class analysis of diagnostic science assessment data using Bayesian networks. Doctoral dissertation, Stanford University, Stanford.
- Steedle, J. T, Shavelson, R. J. (2009) Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, 46(6), 699-715.
- Stigler, S. M. (1982). "Thomas Bayes' Bayesian Inference," *Journal of the Royal Statistical Society, Series A*, 145:250–258.
- Van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer.
- Wandersee, J.H., Mintzes, J.J., & Novak, J.D. (1994). Research on alternative conceptions in science. In D. L. Gabel (Ed.), *Handbook of Research on Science Teaching*. Upper Saddle River, NJ: Merrill/Prentice Hall.
- Wilson, M. (1992) The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, 16, 309-325.
- Wilson, M. (2008) Cognitive diagnosis using item response models. *Journal of Psychology*, 205(2), 74-88.